LEAD SCORING

By: Osama AL-Momani

Problem statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go upas the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

BUSINESS OBJECTIVE

X education wants to know most promising leads.

For that they want to build a Model which identifies the hot leads.

Deployment of the model for the future use.

SOLUTION METHODOLOGY

- **Data cleaning and data manipulation.**
- 1. Check and handle duplicate data.
- 2. Check and handle NA values and missing values.
- 3. Drop columns, if it contains a large number of missing
- values and are not useful for the analysis.
- 4. Imputation of the values, if necessary.
- 5. Check and handle outliers in data.

EDA

- 1. Univariate data analysis: value count, distribution of variables, etc.
- 2. Bivariate data analysis: correlation coefficients and pattern between the
- variables etc.
- 3. Feature Scaling & Dummy variables and encoding of the data.
- 4. Classification technique: logistic regression is used for model making
- and prediction.
- 5. Validation of the model.
- 6. Model presentation.
- 7. Conclusions and recommendations

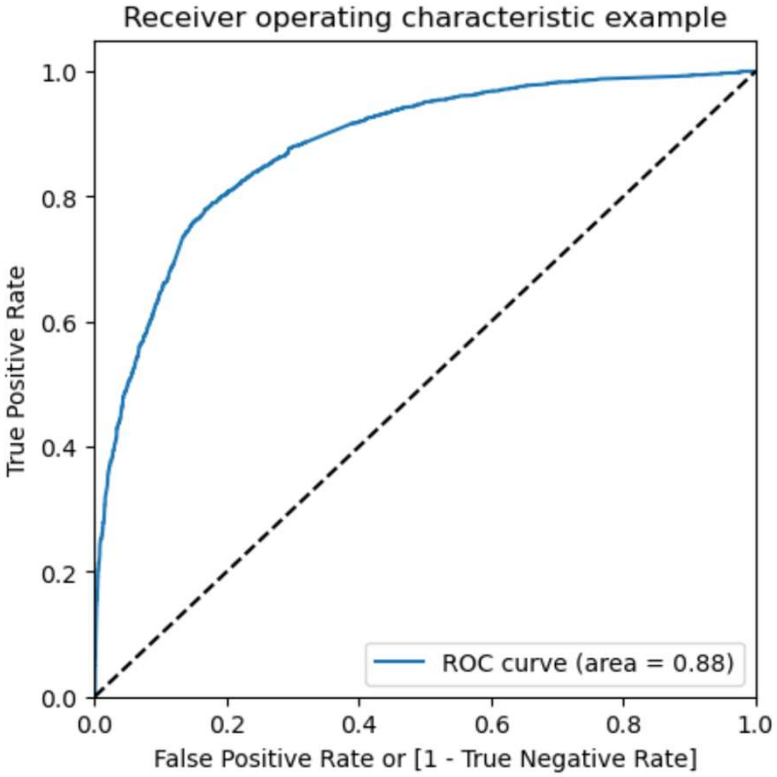
DATA CONVERSION

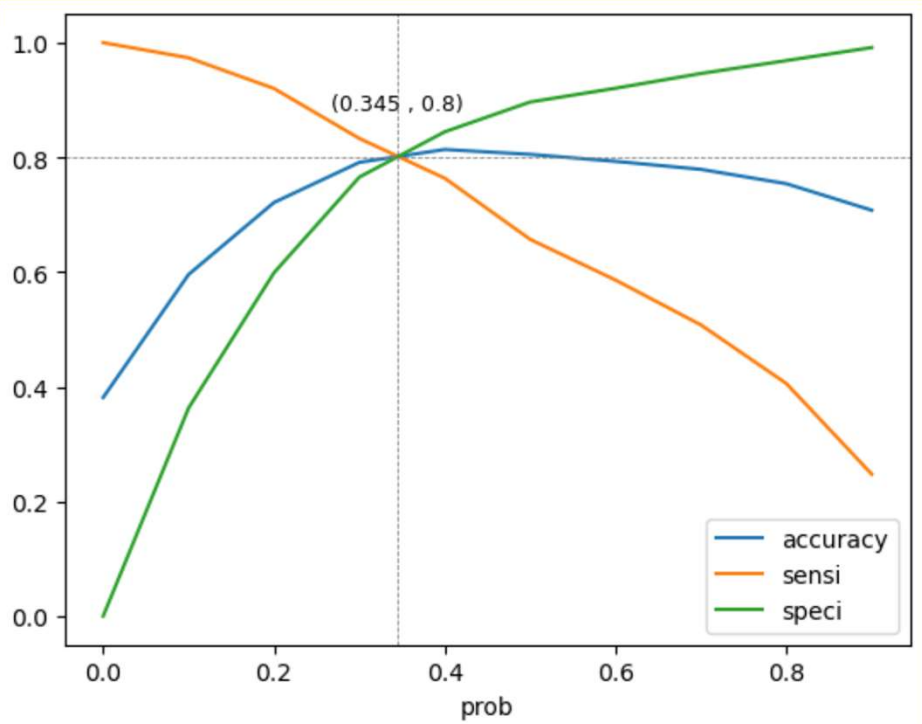
- Numerical Variables are normalized
- Dummy Variables are created for object type variables
- Total Rows : 9240
- Total Columns : 37

MODEL BUILDING

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p-value is greater than 0.05 and vi value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

Curves





Finding Optimal Cut off Point

Optimal cut-off probability is that Probability where we get balanced sensitivity and specificity.

From the second graph it is visible that the optimal cut off is about at 0.35.

CONCLUSION

- The variables of utmost significance among potential buyers were identified as follows (in descending order):
- The total time spent on the website.
- The total number of visits.
- The lead source, particularly:
  - Google
  - Direct traffic
  - Organic search
  - Welingak website
- The last activity, especially:
  - SMS interactions
  - Olark chat conversations
- The lead origin as Lead add format.
- The current occupation indicating "working professional."
- Considering these key factors, X Education stands to thrive by capitalizing on the high likelihood of persuading nearly all potential buyers to reconsider and ultimately invest in their courses.